 Human Molecular Genetics, 1993, Vol. 2, No. 7 1007–1014

Haplotype analysis to determine the position of a mutation among closely linked DNA markers

Michele Ramsay*, Robert Williamson, Xavier Estivill[1], Brandon J.Wainwright+, Meng-Fatt Ho, Stephanie Halford, Juha Kere[2], Erkki Savilahti[3], Albert de la Chapelle[2], Marianne Schwartz[4], Martin Schwartz[5], Maurice Super[5], Peter Farndon[6], Carol Harding[6], Linda Meredith[7], Layla Al-Jader[7], Claude Ferec[8], Mireille Claustres[9], Teresa Casals[1], Virginia Nunes[1], Paolo Gasparini[10], Anna Savoia[10], Pier Franco Pignatti[10], Giuseppe Novelli[11], Massimo Bennarelli[11], Bruno Dallapiccola[11], Luba Kalaydjieva[12] and Peter J. Scambler

Department of Biochemistry and Molecular Genetics, St Mary's Hospital Medical School, London, W2 1PG, UK, [1]Department of Molecular Genetics, Hospital Duran i Reynals, Hospitalet de Llobregat, Barcelona, Spain, [2]Department of Medical Genetics, University of Helsinki, [3]Children's Hospital, University of Helsinki, Finland, [4]Section of Clinical Genetics, University Hospital Rigshospitalet, Kopenhagen, Denmark, [5]Clinical Genetics Unit, Royal Manchester Children's Hospital, Pendlebury, Manchester, [6]Clinical Genetics Unit, Birmingham Maternity Hospital, Edgbaston, Birmingham, [7]Institute of Medical Genetics, University Hospital of Wales, Heath Park, Cardiff, UK [8]Centre de Transfusion Sanguine, Brest, [9]Inserm U 249, Institute de Biologie, Faculte de Medicin, Montpellier, France, [10]Istituto di Scienze Biologiche, Universita de Verona, [11]Department of Public Health and Cell Biology, IInd University of Rome and University of Urbino, Italy and [12] Department of Clinical Genetics, Institute of Obstetrics, Sofia, Bulgaria

Received January 26, 1993; Revised and Accepted April 30, 1993

Positional cloning involves first finding linkage between an inherited phenotype (such as a disease) and a DNA marker, followed by the use of a variety of physical and genetic mapping techniques to move from linkage to mutation. If there is a founder effect within a population, crossovers are often rare between the mutation causing the phenotype and closely situated markers and increasing disequilibrium may be observed as the site of the mutation is approached. Standard coefficients of disequilibrium may, however, be insensitive to the relative position of close markers and the mutation, because they depend upon allele frequencies in the normal population compared to those of the founder chromosome. Using cystic fibrosis in European populations as a model system, alternative methods for determining the position of a mutation are discussed. These include haplotype parsimony and three-way interval likelihood analysis. Both methods predict the location of the major CF mutation accurately from a real set of more than 600 European CF chromosomes.

INTRODUCTION

Haplotype analysis of normal and disease-associated chromosomes provides information for mapping a mutation within tightly linked DNA markers. Linkage disequilibrium has been used to determine how close a particular polymorphic marker is to a disease locus, but it is highly sensitive to the frequency of the polymorphic alleles, and whether the common allele on normal chromosomes is the same as the common allele on the chromosome carrying the mutation (1–3). An attempt is made to explore the use of different methods of haplotype analysis to predict the position of a disease causing mutation within a set of closely linked markers.

The cystic fibrosis (CF) locus was mapped by linkage analysis to chromosome 7q31 (4–7), and it was shown that there is locus homogeneity for this disorder. Alleles of some closely linked markers show significant linkage disequilibrium with the CF mutation in every population tested, particularly in northern Europe. This suggests that the majority of CF chromosomes share a common origin and have the same mutation (8–16). A large haplotype data set was collected for markers close to and within the CF locus on CF and non-CF chromosomes in ten European populations.

In the identification of the CF gene, there were no associated chromosomal anomalies to aid movement from linkage to locus. Therefore, positional cloning was used to identify and characterise the CF transmembrane conductance regulator (CFTR) gene (1, 17). The common disease causing mutation was identified as a 3bp deletion in exon 10, leading to the loss of a phenylalanine residue in the first nucleotide binding fold. This common mutation, ΔF508, is present on between 70% and 80% of CF chromosomes of north European origin, but only 35% to 50% of southern European origin, indicating that there are many other CF causing mutations (18). It is apparent that there is greater mutation homogeneity in northern Europe. In this study, over 600 CF-associated and over 600 non-CF-associated chromosome haplotypes were determined in 10 European populations, using 8 RFLP markers closely linked to the CF mutation. Two of the

*To whom correspondence should be addressed at present address: Department of Human Genetics, South African Institute for Medical Research, PO Box 1038, Johannesburg, 2000, South Africa

+Present address: Centre for Molecular Biology and Biotechnology, University of Queensland, Brisbane, Australia

markers, pG2 and H80, are situated within the CFTR gene, 3' to the common ΔF508 mutation (19).

There are four major prerequisites for the methods of analyses described here: 1. The existence of a single predominant disease-causing mutation. 2. The location of this mutation within a specific haplotype which is referred to as the 'ancestral' haplotype, which is the commonest disease-associated haplotype in all populations studied. 3. Knowledge of the physical order of the markers. 4. The ability to analyze different disease-associated haplotypes which have arisen from the 'ancestral' type through historical recombination.

The analyses identified a region of 225kb, between the markers pMP6d-9 and pG2 as the most likely position for the common CF causing mutation. This is, in fact, the position of the ΔF508 mutation. The occurrence of the ΔF508 mutation within the haplotypes of the present study were not, and are still not known.

The complete compilation of haplotypes is made available in order that others may perform alternative analyses. The analysis of multi-site haplotypes may well offer a paradigm for future studies with other genetically homogeneous disorders.

RESULTS

Allelic association at each of the loci

Allele frequencies for CF and non-CF chromosomes, standardised linkage disequilibrium (delta), odds ratio and chi squared homogeneity and heterogeneity calculations (20) are shown in Table 1 for each of the six RFLPs: pXV-2c/TaqI, pKM.19/PstI, pMP6d-9/MspI, pG2/XbaI, H80/PstI and pJ3.11/MspI. Data for *MET* have been reported previously and are not included in the table. The physical order of the DNA markers and the distances between them are shown in Figure 1, as are the calculated odds ratios for each of the loci and the associated 95% confidence limits. For arithmetic convenience, the alleles were inverted for pKM.19, pMP6d-9, pG2 and H80. A marked increase in allelic association is seen for pKM.19 and pMP6d-9, which are not significantly different from one another, with a decrease in association for the other markers in both the pJ3.11 and *MET* directions. The allele frequencies for the different populations are homogeneous with respect to CF for five of the loci. There is, however, significant ($p < 0.011$) heterogeneity for the H80/PstI alleles.

Three-way analysis of loci

Three-way analysis to determine the most likely position of the major CF mutation was done for each marker-marker combination using the following markers: pXV-2c, pKM.19, pMP6d-9, pG2 and H80. CF and non-CF haplotypes of the paired loci are shown in Table 2 for all the populations combined and then for the northern (Finnish, Danish, both English, Welsh and French) and southern (Spanish, both Italian and Bulgarian) data subsets. These data were used in the calculations which aim to exclude or include the CF mutation from each of the marker-marker intervals. Several assumptions are made, as follows: the predominant CF mutation has occurred only once, on the 'ancestral' CF haplotype; other CF haplotypes have arisen from the 'ancestral' haplotype through recombination; sequential recombination events on the same chromosome are negligible; the frequency of alleles on normal chromosomes has not changed significantly since the CF mutation occurred.

Each marker–marker pair was examined sequentially as is exemplified in the methods section. All the chi squared values which show significant exclusion ($p < 0.001$) of the CF mutation from either within or outside marker–marker intervals are given in Table 3. The results are shown schematically in Figure 2 where the bars of exclusion for CF are proportional in thickness to the size of the chi squared. The chi squared values cannot be added for the intervals because some of the same data have been used to calculate overlapping regions of exclusion. This method of analysis may have more power if more markers were used together.

The northern European populations show greater homogeneity than the southern European populations when analyzed separately. Three-way interval probability analysis for the northern group

Table 1. Distribution of alleles at five polymorphic markers closely linked to the CF mutation in ten European populations

POPULATION	pXV-2c					pKM.19					pMP6d-9					pG2					H80					p25.11				
	CF		non-CF		Δ	CF		non-CF		Δ	CF		non-CF		Δ	CF		non-CF		Δ	CF		non-CF		Δ	CF		non-CF		Δ
	1	2	1	2		1	2	1	2		1	2	1	2		1	2	1	2		1	2	1	2		1	2	1	2	
FINNISH	35	3	18	19	0.48	7	31	22	15	-0.42	5	32	20	13	-0.49	2	36	10	27	-0.30	15	23	13	24	0.04	21	17	12	25	0.23
DANISH	51	7	32	26	0.36	8	50	42	16	-0.59	6	44	29	18	-0.52	2	55	12	42	-0.28	2	56	19	38	-0.39	27	23	7	9	0.09
ENGLISH(M)	72	13	41	49	0.42	11	81	70	22	-0.65	10	76	54	32	-0.53	5	81	20	66	-0.25	nt	nt	nt	nt		19	26	23	21	-0.10
ENGLISH(B)	57	7	39	25	0.32	9	55	46	18	-0.58	7	54	33	2	-0.51	4	62	11	55	-0.17	4	60	12	50	0.02	32	26	17	41	0.26
WELSH	24	4	14	14	0.38	6	22	21	7	-0.54	5	23	19	7	-0.56	4	24	8	20	0.17	nt	nt	nt	nt		14	14	14	14	-
FRENCH	108	17	71	51	0.32	18	108	82	41	-0.53	14	109	62	59	-0.43	9	111	28	90	-0.22	8	46	15	36	-0.18	63	62	56	66	0.05
SPANISH	163	57	107	109	0.25	59	161	149	66	-0.43	42	151	118	71	0.41	10	91	21	81	-0.15	18	63	26	57	0.11	101	113	81	130	0.09
ITALIAN(V)	64	32	40	56	0.25	26	70	68	28	-0.44	12	74	57	29	-0.53	13	75	13	75	-	16	45	12	44	0.06	38	42	33	37	0.004
ITALIAN(U)	64	16	37	41	0.34	19	71	63	26	-0.50	16	56	46	25	-0.43	14	59	21	52	-0.11	nt	nt	nt	nt		24	31	29	28	-0.08
BULGARIAN	34	10	19	19	0.28	10	32	28	7	-0.56	6	35	9	8	-0.40	3	47	6	31	-0.17	7	43	10	26	0.19	5	11	8	8	-0.20
TOTAL	677	166	418	409	0.31	173	681	591	246	-0.51	123	654	447	284	-0.47	66	641	150	539	0.17	70	338	197	275	-0.13	344	367	279	379	0.06
[illegible]																														
[illegible]																														
[illegible]																														

shows that the CF mutation is most probably between pMP6d-9 and pG2. Each of the other intervals is excluded at $p<0.001$. The same region shows the smallest exclusion value for the southern European group, although in this data subset every interval is technically excluded.

Haplotype analysis

Two sets of haplotype data were compiled by hand. The first consists of haplotypes for CF and non-CF chromosomes with the four marker combination pXV-2c, pKM.19, pMP6d-9 and pG2 (Table 4). In total, 547 CF haplotypes were unambiguously assigned. Complete information was obtained for an additional 102 chromosomes but phase could not be deduced. As expected, a higher proportion (0.23) of southern European CF chromosomes were not determinable when compared to the northern group (0.10). The second consists of haplotypes including five markers, the above four and H80, presented in Table 5. In this case a total of 278 CF haplotypes were unambiguously assigned and in a further 70 phase could not be determined. The main reason for the decrease of numbers in the second set is that the Manchester, Welsh and Urbino families were not typed for H80, and in many of the other populations the H80 data were incomplete.

The CF haplotypes were analyzed for compatibility with the proposed major ancestral CF haplotype 1 2 2 2 2. They were compared with regard to the alleles they have in common with the ancestral type. The shaded boxes around the haplotypes highlight the ancestral type or part thereof (Figure 3). The vertical lines show the interval which is most likely to include the CF mutation. There was no apparent difference between the northern and southern groups. When the four locus haplotypes are analyzed, 54 CF chromosomes are incompatible with the CF mutation being between pKM.19 and pMP6d-9 and only two are incompatible with the interval pMP6d-9 to pG2. Similarly, in the five locus analysis 27 and 1 CF chromosomes are incompatible with each of these intervals, indicating that the CF mutation is most likely to be between pMP6d-9 and pG2. Incompatibility refers to a situation where neither of the alleles at markers flanking the proposed position of the CF mutation is the ancestral type.

DISCUSSION

A large body of data on polymorphic markers physically and genetically close to *CF*, have been collected and analyzed with the aim of predicting the position of the CF mutation among them. The order of the markers has been well established by physical mapping using pulse field gel electrophoresis (7,10,19,21,22). Before the cloning of the CF gene, in the absence of chromosomal translocations and deletions, it was difficult to order close markers with respect to the major mutation. When linked markers are at a genetically determinable distance from the mutation, it is possible to use information from recombinant families, but when the distance becomes so small that recombinant families are very rare, this is no longer possible. In such cases alternative methods are needed. An attempt has been made to analyze haplotypes constructed from closely linked markers in order to determine the position of the major CF mutation.

Three different methods of analysis were used. The first is the widely used method of allelic association (20). Figure 1 shows that alleles at pKM.19 and pMP6d-9 clearly have the highest association with CF, indicating that they are likely to be the closest markers to the mutation. They are not significantly different from one another and theoretically the CF mutation could lie between,

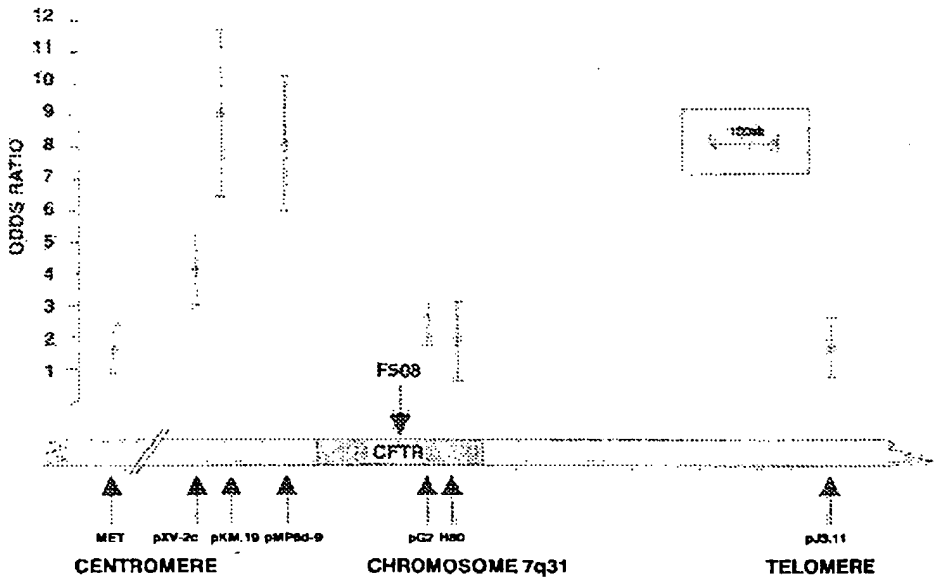

Figure 1. Physical distance and relative positions of markers closely linked and flanking the major CF mutation, ΔF508, are shown. Scale is indicated in the box, top right. The vertical axis shows the odds ratio and 95% confidence intervals as a measure of allelic association for each of the polymorphic markers with CF. The value for *MET* was calculated from the data for the TaqI/metH RFLP as reported in reference 11.

Table 2. Pairwise analysis of all the marker/marker combinations. Haplotype numbers and standardised linkage disequilibrium values are given for the combined sample (TOTAL) then separately for the north and south European populations on CF and non-CF chromosomes

HAPLO	TOTAL		NORTH		SOUTH	
	CF	N	CF	N	CF	N
TOTAL	814	765	386	394	428	371
Δ	-0.37	-0.05	-0.41	-0.04	-0.34	-0.06
TOTAL	750	673	366	359	384	314
Δ	-0.29	-0.25	-0.46	-0.32	-0.15	-0.17
TOTAL	691	645	382	382	309	263
Δ	-0.10	-0.06	-0.14	-0.08	-0.06	-0.04
TOTAL	387	351	213	206	174	145
Δ	-0.09	0.05	-0.11	0.07	-0.07	0.01
TOTAL	735	676	360	354	375	322
Δ	0.68	0.64	0.76	0.60	0.52	0.70

HAPLO	TOTAL		NORTH		SOUTH	
	CF	N	CF	N	CF	N
TOTAL	711	641	384	378	327	263
Δ	-0.10	-0.40	-0.06	-0.45	-0.14	-0.33
TOTAL	397	331	209	196	188	135
Δ	-0.11	-0.50	-0.06	-0.48	-0.17	-0.53
TOTAL	669	594	366	354	303	240
Δ	-0.06	-0.25	-0.01	-0.25	-0.11	-0.24
TOTAL	368	293	199	178	169	115
Δ	-0.10	-0.53	-0.08	-0.55	-0.13	-0.47
TOTAL	388	330	216	202	172	128
Δ	0.50	0.64	0.39	0.62	0.58	0.68

CF = CF-associated chromosomes
N = non-CF chromosomes.

Table 3. Pairwise analysis of markers closely linked to the CF mutation. Chi square values given where the CF mutation is excluded from physical intervals at $p < 0.001$

HYPOTHESIS				TOTAL	NORTH	SOUTH
				55.68	ns	
				64.07		
				14.28	ns	25.31
				200.26	113.73	80.19
				202.29	101.74	71.02
				27.40	ns	28.49
CF	X	KM	G2			
KM	G2	X	CF			

ns — not excluded at $p < 0.001$.
X — position of the recombination event.
XV, pXV-2c; KM, pKM.19; D9, pMP6d-9; G2, pG2.

or to either side but close to, these markers. There is a marked decrease in association for markers towards both MET and pJ3.11. The two loci, pG2 and H80, show significantly lower and non-overlapping odds ratio values when compared to pKM.19 and pMP6d-9, indicating that the CF mutation is unlikely to lie to the pJ3.11 side of them. There is a tendency for delta values to be slightly lower in the southern European populations for the markers pXV-2c, pKM.19 and pMP6d-9, indicating more heterogeneity than for the northern group, as reported previously (8, 9).

The analysis of standardised linkage disequilibrium coefficients on CF and non-CF chromosomes showed that the values were not consistently higher on CF chromosomes as opposed to non-CF chromosomes as might be expected. The pairwise comparisons where delta values were higher on non-CF chromosomes were: pKM.19, pG2; pKM.19, H80; pMP6d-9, pG2; pMP6d-9, H80; and pG2, H80. Each of these comparisons includes either pG2 or H80 and an explanation may be that a second CF mutation has occurred on chromosomes where either pG2 and/or H80 is allele 1. Seven percent of CF chromosomes in the north fall within this group, and 18% in the south.

The other two analytical methods assume that the CF mutation occurred only once on an 'ancestral' CF haplotype, and any other CF haplotypes have arisen from it through recombination with non-CF chromosomes, sequential recombination events are rare and the RFLPs which are studied predate the CF mutation.

In the north European subset 81% of the CF chromosomes are haplotype 1 2 2 2 (pXV-2c, pKM.19, pMP6d-9, pG2). This haplotype is present on 63% of southern European CF chromosomes indicating that this subset is less homogeneous. It is therefore reasonable to assume, at least in the northern group, that there is a single predominant mutation which causes CF, and other mutations at this locus make a small contribution to the total data. It is now known that the major CF mutation, ΔF508, is present on 70 to 80% of CF chromosomes in north European populations and 35 to 45% of south European populations (18).

There is some evidence suggesting that groups of patients with severe symptoms have a more homogeneous haplotype composition whereas patients with milder symptoms have several different haplotypes (17). It would have been most suitable to analyze the data from a clinically homogeneous group of CF patients, but data on severity are not available for most of the cohorts presented here. If a second ancestral chromosome had made a large contribution to the haplotypes in our data set this may be apparent as a second common haplotype associated with CF. Haplotype data with the present markers are not available in patients assigned to PI or PS groups, but a previous publication has indicated that a high proportion of PS cases are allele 1 at a locus identical to pKM.19 (17). Recent studies on Italian, Belgium and German CF families showed that although there is a strong correlation between the ΔF508 mutation and the

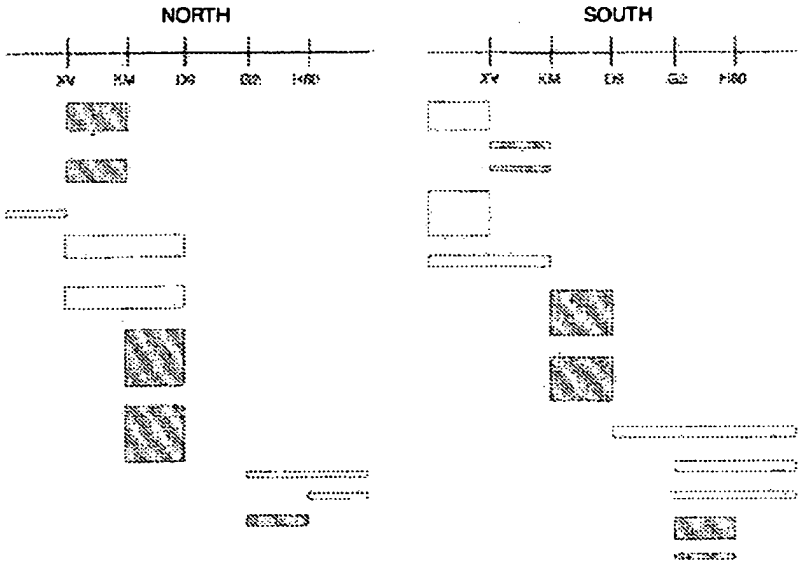

Figure 2. Schematic representation of the three-way exclusion analysis of the haplotype data. Chi squared values for exclusion at $p<0.001$ are shown as blocks between the markers and their height is proportional to their chi squared values. The data for the northern and southern European populations are analysed separately. The boxes are shaded when exclusion data was generated from adjacent markers. XV = pXV-2c; KM = pKM.19; D9 = pMP6d-9; G2 = pG2.

Table 4. Distribution of CF- and non-CF-associated (CF and N) haplotypes (pXV-2c, pKM.19, pMP6d-9, pG2) in European populations

HAPL	FINNISH		DANISH		ENG(M)		ENG(B)		WELSH		FRENCH		SPANISH		ITAL(V)		ITAL(U)		BULG		TOTAL		NORTH		SOUTH	
	CF	N	CF	N	CF	N	CF	N	CF	N	CF	N	CF	N	CF	N	CF	N	CF	N	CF	N	CF	N	CF	N
1 2 2 2	26	3	35	3	61	4	43	2	17	1	87	2	53	6	32	6	29	2	21	1	404	30	269	15	135	15
1 2 2 1	0	4	1	3	2	3	1	4	2	1	5	10	1	4	5	3	5	4	0	0	22	36	11	25	11	11
1 1 2 2	1	2	2	3	1	8	1	6	2	3	5	18	2	4	3	5	0	3	2	0	19	52	12	40	7	12
1 1 1 2	3	2	1	14	4	18	1	12	1	4	1	24	8	25	1	13	1	6	0	3	21	121	11	74	10	47
2 2 2 2	0	0	1	0	4	2	0	3	0	0	2	6	2	5	7	3	3	4	4	1	23	24	7	11	16	13
2 2 2 1	0	1	0	3	1	7	0	2	0	1	2	7	5	7	4	2	0	3	0	1	12	34	3	21	9	13
2 1 1 2	0	12	1	11	5	27	0	12	2	7	10	26	10	30	4	29	1	12	1	3	32	169	16	95	16	74
2 1 2 2	0	0	0	0	0	1	2	0	0	1	0	2	2	1	4	2	0	1	0	0	8	8	2	4	6	4
1 2 1 2	0	0	0	0	0	0	1	1	1	2	0	1	0	0	0	1	2	0	0	0	4	5	2	4	2	1
1 2 1 1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	1	0	1	1	0
2 1 1 1	0	2	0	1	0	4	0	2	1	5	0	4	0	4	0	3	0	1	0	0	1	26	1	18	0	8
1 1 1 1	0	1	0	1	0	1	0	0	0	0	0	2	0	2	0	0	0	1	0	0	0	8	0	5	0	3
2 1 2 1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
2 2 1 2	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	2	0	2	0	0
2 1 2 1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	2	0	0	0	2
TOTAL	30	27	41	40	76	75	49	45	26	25	112	104	83	89	60	68	42	37	28	9	547	519	334	316	213	203
not det	8	6	8	4	6	6	8	6	2	2	6	4	16	12	24	16	16	20	8	6	102	82	38	28	64	54
incomplete	0	5	9	13	10	10	9	15	0	0	5	10	120	115	12	12	32	32	14	35	214	247	36	53	178	194

ENG(M) — Manchester; ENG(B) — Birmingham; ITAL(V) — Verona; ITAL(U) — Urbino; not det — not determined.

'ancestral' haplotype, this haplotype has also been found in association with other CF mutations (23–25). In the Italian cohort of 212 CF families 27% of CF-associated chromosomes with the 1 2 2 (pXV-2c pKM.19 pMP6d-9) haplotype had a non-ΔF508 mutation (23).

When analyzing haplotype data, a certain number of chromosomes are lost to analysis because of incomplete data, and of haplotypes for which phase cannot be determined because both parents and all available children are heterozygous at one or more loci. The second factor is of greater concern because it is not random, and potentially lowers the proportion of CF haplotypes which are not the 'ancestral' type, because inability to determine phase is often due to the occurrence of the 'ancestral' CF haplotype in conjunction with a 'rare' CF haplotype. This effect will be less marked in the set of non-CF chromosomes, as there is a more widespread distribution of haplotypes, and loss of information is more likely to be random.

In the three-way analysis of the data the hypotheses were tested that the CF mutation is between or outside each marker–marker interval. Chi squared values were determined to examine the likelihood of the CF mutation for each interval, and the position was rejected when a significance level of $p<0.001$ was obtained. The schematic representation of the data in Figure 2 shows that in the northern group, there was no exclusion in the interval pG2 to pMP6d-9, indicating that this is the most likely position for the CF mutation. In the southern subset, this interval showed the lowest level of exclusion. The interval pKM.19 to pMP6d-9 was consistently excluded with high chi squared values.

Table 5. Distribution of CF- and non-CF-associated (CF and N) haplotypes (pXV-2c, pKM.19, pMP6d-9, pG2, H80) in European populations

HAPLOTYPES	FINNISH		DANISH		ENGLISH		FRENCH		SPANISH		ITALIAN		BELG		TOTAL		NORTH		SOUTH	
	CF	N	CF	N	CF	N	CF	N	CF	N	CF	N	CF	N	CF	N	CF	N	CF	N
1 2 2 2 2	32	1	35	2	35	3	33	1	32	2	34	2	26	0	369	69	126	5	66	5
2 1 1 2 2	32	2	0	0	0	1	3	1	3	1	3	2	3	1	45	8	15	4	7	4
1 1 1 2 2	0	0	0	0	0	0	1	3	0	1	0	0	0	0	1	5	1	2	0	3
1 2 2 1 1	0	1	1	2	1	1	1	2	1	3	4	6	0	0	8	18	3	15	0	3
1 1 2 2 2	1	2	0	1	1	0	2	7	2	2	2	2	1	0	11	27	6	17	2	4
2 1 2 2 2	5	2	1	14	2	13	3	11	0	0	1	2	0	1	11	53	6	40	3	17
2 2 2 2 2	0	0	1	0	0	0	0	3	1	0	0	0	1	0	3	3	2	0	1	3
2 2 1 1 2	0	0	0	0	0	0	2	0	0	0	0	0	0	0	2	0	2	0	0	0
2 1 1 2 2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0
1 2 2 2 1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2 1 1 2 1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2 2 2 2 2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1 2 2 1 2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2 2 2 1 2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1 2 2 1 1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2 1 2 1 2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2 1 2 1 1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1 2 1 2 1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2 1 1 2 1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2 1 1 1 1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TOTAL	70	71	43	32	43	42	49	43	41	44	43	32	28	5	438	226	167	148	81	33
[illegible]	2	28	8	6	19	6	4	6	12	2	29	15	8	2	80	65	22	18	40	18
[illegible]	68	3	9	13	9	15	9	5	165	179	46	48	16	58	337	249	18	31	219	267

Abbreviations as for Table 4.

The analysis of the markers pKM.19 and pMP6d-9 is of interest, as it shows the greatest difference between the northern and southern data sets. The haplotypes 1 2 and 2 1 (pKM.19, pMP6d-9) are over-represented on CF chromosomes in the southern subset, which can be explained either by an early recombination event between these markers, or by further CF mutations which occurred in the south on those specific haplotypes. Neither of the marker orders, CF pKM.19 pMP6d-9 nor pKM.19 pMP6d-9 CF, are excluded in the north European data set, suggesting that the two markers are in strong disequilibrium. The order, CF pKM.19 pMP6d-9, is excluded with a lower chi squared value than the correct locus order, pKM.19 pMP6d-9 CF, though it may not be significantly different.

The haplotypes were also analyzed using the parsimony method to determine the position of the CF mutation as predicted by the least number of crossover events. The most common CF haplotype is 1 2 2 2, 81% in the north and 63% in the south. The next most common haplotypes on CF chromosomes are 1 1 1 2, 2 1 1 2 and 2 2 2 2. The first two could have arisen through recombination between pMP6d-9 and pG2 ($\Delta = -0.25$ on non-CF chromosomes) and the third by recombination between pXV-2c and pKM.19 ($\Delta = -0.05$ on non-CF chromosomes). Delta values for the other intervals, pKM.19 to pMP6d-9 and pG2 to H80, are both 0.64 on non-CF chromosomes and it is therefore less likely that recombination will occur between them. It is thus reasonable to assume that the three most common haplotypes shown above are derived from the ancestral CF chromosome by single historical crossover events. The schematic representation of the data in Figure 3 shows that the interval pMP6d-9 to pG2 is the most likely position of the CF mutation. The interpretation of the data may well be influenced by the relative lack of informativity of the pG2 and H80 RFLPs, since a proportion of recombinant events between them and adjacent markers will not be visible. This is the reason for not quantitating the observations from the CF haplotypes by computing the expected numbers of each recombination event.

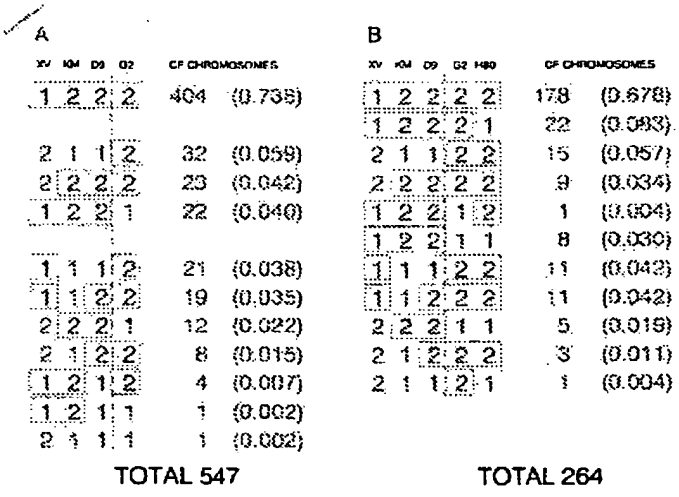

Figure 3. CF-associated haplotypes for the four(A) and five(B) marker haplotype data in the combined data set. The ancestral CF haplotype and portions thereof have been placed in shaded boxes. The dashed vertical line shows the region of maximum overlap which is the most likely position of the CF mutation. XV = pXV-2c; KM = pKM.19; D9 = pMP6d-9; G2 = pG2.

We have attempted to present the data in full, in order that others may use different methods of analysis to position the CF mutation relative to the DNA markers. The allelic association calculations indicate that the major CF mutation should be close to the markers pKM.19 and pMP6d-9. Both other analytical methods argue for the location of the CF mutation between pMP6d-9 and pG2, an interval of 225kb. This is most convincingly shown for the more homogeneous north European cohort, using the three-way interval probability analysis and (for the total data set) haplotype by inspection. The ΔF508 mutation is, in fact, situated within this 225kb region showing that these methods accurately predict its location. The methods presented

here should be applicable to other systems where closely linked polymorphic markers cannot be ordered with respect to a mutation.

MATERIAL AND METHODS

CF families

CF families from the following eleven centres in Europe were included in the present study: Finland (10 families); Denmark (29); England, Manchester (46); England, Birmingham (33); Wales, Cardiff (14); France, Paris (58); France, Montpellier (6); Spain (110); Italy, Verona (48); Italy, Urbino (45) and Bulgaria (25). In total, 433 nuclear families with at least one living affected child were studied to deduce the phase of the haplotypes. The families are not known to be related. Each of the groups has been analyzed separately with the exception of the French cohorts where they were pooled because of the small sample from Montpellier. Subsets of the data from several of the populations were reported previously (7, 9, 10, 26–30).

Population description and clinical profile

The incidence of CF is similar in most European populations, with a carrier frequency of 0.04 to 0.05, with the exception of Finland where it is considerably lower, at 0.013. CF diagnostic criteria included at least one positive sweat test and pulmonary and/or pancreatic disease. In Finland and Denmark CF is clinically very homogeneous and all but one CF family from each group have pancreatic insufficiency (PI); the English and Welsh CF families are also mostly PI. The French cohort are mostly of Celtic origin, and reside in north west Brittany; they are also predominantly PI and were included in the northern European data subset. The Italian CF families from Verona come from the Veneto region of north east Italy and those from Urbino are mainly from central Italy, and show a higher proportion (0.15 to 0.20) of pancreatic sufficient (PS) families than in northern Europe. The Spanish CF families are from all over the country and it is not known what proportion are PI; the Italian and Spanish families were designated to the southern European subset (29). No clinical data was available for the Bulgarian CF families but, they showed a similar genetic profile to other southern European groups and were included in this subset.

RFLPs linked to the CF locus

The RFLPs included the known CF flanking markers, pXV-2c and pJ3.11, and several markers between them. Each of the markers has been previously described: pXV-2c and pKM.19 (7,8), pMP6d-9 (10), pG2 and H80 (19) and pJ3.11 (5). pXV-2c, pKM.19, pMP6d-9, pG2, H80 and pJ3.11 were used to detect TaqI, PstI, MspI, XhaI, PstI and MspI RFLPs, respectively. Probe inserts were radiolabelled by random oligonucleotide primed DNA synthesis (31). Family DNA was prepared from peripheral blood by standard methods. Genomic DNA was digested with the appropriate enzyme using conditions recommended by the manufacturer, size fractionated on agarose gels, capillary transferred to nylon membranes, hybridised and autoradiographed.

Analysis of data

Three methods of analysis were used.

1. Allelic association at each of the loci. Allele frequencies at each locus were determined using the gene counting method. Standardised linkage disequilibrium, relative risk, chi squared heterogeneity and homogeneity, and odds ratio calculations were described previously (8, 9, 20).

2. Three-way analysis of loci. Haplotypes were assigned by hand from genotypings on nuclear CF families. A computer program, HAPLOT, was written for the purpose of calculating chi squared and G values for each of the possible positions of the CF locus with respect to each pairwise combination of markers. The log likelihood ratio, G, was used in preference to the chi squared when any observed or expected value was less than 5. G is defined by the relation: $G = 2^{*}(fobs^{*}\ln(fobs/fexp))$ and its distribution approximates to that of the chi squared statistic (32). The algorithm of the analysis is explained with worked examples below. Since HAPLOT examines marker pairs within haplotypes it is possible to include some of the data from the incomplete haplotypes and from haplotypes where phase could not be determined.

The proposed major 'ancestral' CF haplotype is 1 2 2 2 2 (pXV-2c, pKM.19, pMP6d-9, pG2, H80) and only the CF haplotypes which differ from it, ie. the recombinant haplotypes, are analyzed. The data were analyzed as a combined set and divided into the northern and southern European data sets separately. As an example, the markers pXV-2c and pKM.19 are analyzed with respect to the most likely position of the CF mutation. One examines the likelihood of the following locus orders: CF pXV-2c pKM.19; pXV-2c CF pKM.19; and pXV-2c pKM.19 CF. It has to be borne in mind that the locus order pXV-2c CF pKM.19 can have recombinations on CF chromosomes in two possible positions: pXV-2c × CF pKM.19 and pXV-2c CF × pKM.19 (× = position of crossover). In order to maximise the power of the three-way analysis, the locus1 × CF locus2 and locus1 CF × locus2 analyses which test the same locus order should, in some way, be combined.

In the analysis of the combined data set, the observed number of CF × pXV-2c recombinants is the number of CF chromosomes with pXV-2c allele 2, since the 'ancestral' haplotype has allele 1. There are 64 CF-associated 2 1 haplotypes (pXV-2c, pKM.19) and 73 CF-associated 2 2 haplotypes, making the total number of recombinant chromosomes studied in this analysis 137. Since this type of recombination event would place alleles originally found on non-CF chromosomes adjacent to the CF mutation, we would expect to find haplotypes 2 1 and 2 2 at a frequency similar to that found on non-CF chromosomes, assuming the locus order to be CF pXV-2c pKM.19. By counting the numbers of 2 2 and 2 1 haplotypes on non-CF chromosomes we can calculate their frequencies and obtain the expected values for the recombination chromosomes by multiplying the frequency by the total number of observed recombinant chromosomes. In this case the frequency of the non-CF-associated haplotype 2 1 is 0.75 (278/370) and that of haplotype 2 2 is 0.25 (92/370). The *expected* number of CF chromosomes with haplotype 2 1 is therefore 102.94 (0.75 × 137) and those with haplotype 2 2 is 34.06 (0.25 × 137). To assess this difference statistically we used the chi squared test. For this example we found the chi squared to be 55.68 ($p < 0.001$), and therefore it is unlikely that the CF mutation lies toward the centromere from pXV-2c.

Using the same markers, the next hypothesis tested is that the CF mutation lies between the markers pXV-2c and pKM.19 and that the recombinants have occurred between CF and pXV-2c. In this case, the recombinant from the 'ancestral' haplotype would be CF chromosomes containing the pXV-2c allele 2. Only haplotype 2 2 would be a single recombinant since haplotype 2 1 would imply a double recombination event from the major 'ancestral' haplotype. Double recombinants with haplotype 2 1 would be expected to be very rare. The expected number of chromosomes with haplotype 2 2 is calculated using the frequency of pXV-2c allele 2 on non-CF chromosomes and is 65.76 (0.48 × 137). The chi squared is 55.40 ($p < 0.001$) indicating that it is unlikely that CF lies between these markers.

The next hypothesis tested is that the CF mutation lies between the markers pXV-2c and pKM.19 and that the recombinants have occurred between CF and pKM.19. In this case, the recombinant from the 'ancestral' haplotype would be CF chromosomes containing the pKM.19 allele 1. Only haplotype 1 1 would be a single recombinant since haplotype 2 1 would imply a double recombination event from the major 'ancestral' haplotype. The expected number of chromosomes with haplotype 1 1 is calculated using the frequency of pKM.19 allele 1 on non-CF chromosomes and is 95.04 (0.72 × 132). The chi squared is 64.67 ($p < 0.001$) confirming that it is unlikely that CF lies between these markers.

The final hypothesis is that the CF locus lies telomeric to both pXV-2c and pKM.19. Recombinant CF chromosomes will have pKM.19 allele 1 and therefore one considers the frequencies of non-CF-associated haplotypes 1 1 and 2 1 (pXV-2c and pKM.19). The *expected* number of 1 1 haplotypes is 66 (0.5 × 132) and that for haplotype 2 1 is also 66 (0.5 × 132). The chi squared is 0.06 ($p = 0.806$) indicating that the CF mutation is most likely to be telomeric to pKM.19 since it has been more significantly excluded from each of the other intervals.

This approach does, however, have several deficiences and should be regarded as an attempt to formalise the three-way analysis, a method upon which others will doubtless improve.

3. Haplotype analysis by inspection.

ACKNOWLEDGEMENTS

We thank the many clinicians and supporting staff for their dedication and the CF families for providing blood samples. This work was funded by the Cystic Fibrosis Research Trust and the U.K. Medical Research Council. M.R. is supported by the South African Medical Research Council and the South African Institute for Medical Research. C.F. acknowledges the help of the AFLM (Association Francaise du Lutte contre la Mucoviscidose) and the ABERM (Association Bretonne d'Etudes et de recherches de la Mucoviscidose). X.E., T.C. and V.M. were supported by the 'Fondo de Investigaciones de la Seguridad Social' (89/0563) and the 'Direccion General de Investigacion Cientifica y Tecnica' (PB87/0074). The work contributed by G.N., M.B., and B.D. was supported by grants from 'The Golden Products of Italy' and by Regione Marche, Italy. P.F.P., P.G. and A.S. thank G. Mastella, Director of the Cystic Fibrosis Center of Verona, for his help and encouragement; the latter two were recipients of fellowships from the CF Centre of Verona. The study of the Finnish families was supported by the Academy of Finland, the Sigrid Juselius Foundation, and the Folkhalsan Institute of Genetics.

REFERENCES

1. Rommens,J.M., Ianuzzi,M.C., Kerem,B-S., Drumm,M.L., Melmer,G., Dean,M., Rozmahel,R., Cole,J.L., Kennedy,D., Hidaka,N., Zsiga,M., Buchwald,M., Riordan,J.R., Tsui,L.-C. and Collins,F.S. (1989) *Science* 245, 1059–1065.
2. Pritchard,C., Cox,D.R. and Myers,R.M. (1991) *Am. J. Hum. Genet.* 49, 1–6.
3. Kaplan,N. and Weir,B.S. (1992) *Am. J. Hum. Genet.* 51, 333–343.
4. Tsui,L-C., Buchwald,M., Barker,D., Braman,J.C., Knowlton,R., Schumm,J.W., Eiberg,H., Mohr,J., Kennedy,D., Plavsic,N., Zsiga,M., Markiewicz,D., Akots,G., Brown,V., Helms,C., Gravius,T., Parker,C., Rediker,K. and Donis-Keller,H. (1985) *Science* 230, 1054–1057.
5. Wainwright,B.J., Scambler,P.J., Schmidtke,J., Watson,E.A., Law,H-Y., Farrall,M., Cooke,H.J., Eiberg,H. and Williamson,R. (1985) *Nature* 318, 384–385.
6. White,R., Woodward,S., Leppert,M., O'Connell,P., Hoff,M., Herbst,J., Lalouel,J-M., Dean,M. and Vande Woude,G. (1985) *Nature* 318, 382–384.
7. Estivill,X., Farrall,M., Scambler,P.J., Bell,G.M., Hawley,K.M.F., Lench,N.J., Bates,G.P., Kruyer,H.C., Frederick,P.A., Stanier,P., Watson,E.K., Williamson,R. and Wainwright,B.J. (1987) *Nature* 326, 840–845.
8. Estivill,X., Scambler,P.J., Wainwright,B.J., Hawley,K., Federick,P., Schwartz,M., Baiget,M., Kere,J., Williamson,R. and Farrall,M. (1987) *Genomics* 1, 257–263.
9. Estivill,X., Farrall,M., Williamson,R., Ferrari,M., Seia,M., Giunta,A.M., Novelli,G., Potenza,L., Dallapiccolla,B., Borgo,G., Gasparini,P., Pignatti,P.F., De Benedetti,L., Vitale,E., Devoto,M. and Romeo,G. (1988) *Am. J. Hum. Genet.* 43, 23–28.
10. Estivill,X., Gasparini,P., Novelli,G., Casals,T., Nunes,V., Gallano,P., Savoia,A., Ruzzo,A., Dallapiccola,B. and Pignatti,P.F. (1989) *Hum. Genet.* 83, 175–178.
11. Schmidtke,J., Krawczak,M., Schwartz,M., Alkan,M., Bonduelle,M., Buhler,E., Chemke,M., Darnedde,T., Domagk,J., Engel,W., Frey,D., Fryburg,K., Halley,D., Hundrieser,J., Ladanyi,L., Libaers,I., Lissens,W., Machler,M., Malik,N.J., Morreau,J., Neubauer,V., Oostra,B., Pape,B., Poncin,J.E., Schinzel,A., Simon,P., Trefz,F.K., Tummler,B., Vassart,G. and Voss,R. (1987) *Hum. Genet.* 76, 337–343.
12. Beaudet,A.L., Feldman,G.L., Fernbach,S.D., Buffone,G.J. and O'Brien,W.E. (1989) *Am. J. Hum. Genet.* 44, 319–326.
13. Kerem,B., Buchanan,J.A., Durie,P., Corey,M.L., Levison,H., Rommens,J.M., Buchwald,M. and Tsui,L-C. (1989) *Am. J. Hum. Genet.* 44, 827–834.
14. Cutting,G.R., Antonarakis,S.E., Buetow,K.H., Kasch,L.M., Rosenstein,B.J. and Kazazian,H.H. (1989) *Am. J. Hum. Genet.* 44, 307–318.
15. Tsui,L-C. (1989) *Am. J. Hum. Genet.* 44, 303–306.
16. Vidaud,M., Kitzis,A., Ferec,C., Bozon,D., Dumur,V., Giraud,G., David,F., Pascal,O., Auvinet,M., Morel,Y., Andre,J., Chomel,J.C., Saleun,J.P., Farriaux,J.P., Roussel,P., Labbe,A., Dastugue,B., Lucotte,G., Monnier,N., Foucaud,P., Goosens,M., Feingold,J. and Kaplan,J.C. (1989) *Hum. Genet.* 81, 183–184.
17. Kerem,B-S., Rommens,J.M., Buchanan,J.A., Markiewics,D., Cox,T.K., Chakravarti,A., Buchwald,M. and Tsui,L.C. (1989) *Science* 245, 1073–1080.
18. The Cystic Fibrosis Genetic Analysis Consortium. (1990) *Am. J. Hum. Genet.* 47, 354–359.
19. Ramsay,M., Wainwright,B.J., Farrall,M., Estivill,X., Sutherland,H., Ho,M-F., Davies,R., Halford,S., Tata,F., Wicking,C., Lench,N., Bauer,I., Ferec,C., Farndon,P., Kruyer,H., Stanier,P., Williamson,R. and Scambler,P.J. (1990) *Genomics* 6, 39–47.
20. Hill,W.G. and Robertson,A. (1968) *Theor. Appl. Genet.* 226–231.
21. Poustka,A., Lehrach,H., Williamson,R. and Bates,G. (1988) *Genomics* 2, 337–345.
22. Collins,F.S., Drumm,M.L., Cole,J.L., Lockwood,W.K., Vande Woude,G.F. and Ianuzzi,M.C. (1987) *Science* 235, 1046–1049.
23. Novelli,G., Gasparini,P., Savoia,A., Pignatti,P.F., Sangiuolo,F. and Dallapiccola,B. (1990) *Hum. Genet.* 85, 420–421.
24. Cuppens,H., Legius,E., Cabello,P., Marynen,P., De Boeck,C., Decorte,R., Fryns,J-P., Eggermont,E., Van den Berghe,H. and Cassiman,J-J. (1992) *Hum. Genet.* 88, 639–641.
25. Dork,T., Neumann,T., Wulbrand,U., Wulf,B., Kalin,N., Maas,G., Krawczak,M., Guillermit,H., Ferec,C., Horn,G., Klinger,K., Kerem,B-S., Zielenski,J., Tsui,L-C. and Tummler,B. (1992) *Hum. Genet.* 88, 417–425.
26. Kere,J., Norio,R., Savilahti,E., Estivill,X., de la Chapelle,A. (1989) *Hum. Genet.* 83, 20–25.
27. Ivinson,A.J., Read,A.P., Harris,R., Super,M., Schwarz,M., Clayton Smith,J. and Elles,R. (1989) *J. Med. Genet.* 26, 426–430.
28. Gasparini,P., Novelli,G., Estivill,X., Oliveieri,D., Savoia,A., Ruzzo,A., Nunes,V., Borgo,G., Antonelli,M., Williamson,R., Pignatti,P.F. and Dallapiccola,B. (1990) *J. Med. Genet.* 27, 17–20.
29. Estivill,X., McLean,C., Nunes,V., Casals,T., Gallano,P., Scambler,P. and Williamson,R. (1989) *Am. J. Hum. Genet.* 44, 704–710.
30. Ferrari,M., Antonelli,M., Bellini,F., Borgo,G., Castiglione,O., Curcio,L., Dallapiccola,B., Devoto,M., Estivill,X., Gasparini,P., Guinta,A., Marianelli,L., Mastella,G., Novelli,G., Pignati,P., Romano,L., Romeo,G., Seia,M. and Williamson,R. (1990) *Hum. Genet.* 84, 435–438.
31. Eriberg,A. und Vogelstein,B. (1984) *Anal. Biochem.* 137, 266–267.
32. Sokal,R.R. and Rohlf,G. (1983) Bionetry 2nd Edition. Freeman, San Francisco.